REMARKS

Claims 1-116 were pending in the application. Claims 1-94 and 97-116 were withdrawn from consideration as directed to non-elected inventions. Claim 95 has been amended to recite the composition of the polypeptide to which the antibody binds. Claim 117 has been added. Support for claim 117 can be found throughout the specification and, for example, at page 51. The title has been amended as requested by the Office. Table 5 has been amended to change "SEQ ID NO." to "SEQ ID NO:" as requested by the Office. The specification has been amended to remove an embedded hyperlink. The specification has also been amended to correct trademark usage.

Upon entry of this amendment claims 95, 96, and 117 will be pending.

No new matter has been added.

Oath/Declaration

According to the Office the Declaration is allegedly defective because it does not identify the citizenship of inventor Roberds.

Applicants will submit a supplemental declaration indicating the citizenship of Roberds under separate cover.

Specification

The Office alleges that the title is not descriptive. Applicants respectfully disagree. However, in order to further prosecution, Applicants have amended the title.

The Office objected to the use of "SEQ ID NO." in Table 5 of the present specification. In order to further prosecution, Table 5 has been replaced with an identical table except "SEQ ID NO." has been replaced with "SEQ ID NO:", rendering the objection moot.

The specification was objected to for allegedly improper trademark and hyperlink usage. Applicants have amended the specification so that the trademarks are identified properly and the embedded hyperlink has been removed. Applicants note that the

trademarks used on page 24 are for plasmids and should begin with lowercase letters because this is the standard usage in the art.

In view of the foregoing, Applicants respectfully request that the objections to the specification be withdrawn.

Objections

Claims 95 and 96 stand objected for allegedly being of improper dependent form for failing to further limit the subject matter of a previous claim. Claim 95 has been amended so that it no longer depends on a withdrawn claim and has been rewritten in independent form. Claim 96 properly depends from claim 95.

In view of the foregoing, Applicants respectfully request that the objections be withdrawn.

Rejection under 35 U.S.C. § 101

Claims 95 and 96 stand rejected under 35 U.S.C. § 101 because the claimed invention is allegedly not supported by a specific, substantial and credible asserted utility or a well established utility. The Office also alleges that "the instant application has provided a description of an isolated DNA encoding a protein and the protein encoded thereby. Because the instant application does not disclose the biological role of this protein or its significance, an antibody to the protein cannot be considered particularly useful." (Office Action, page 5). Applicants respectfully disagree.

Utility Examination Guidelines

The Utility Examination Guidelines require a claimed invention have a specific, substantial and credible asserted utility, or, alternatively a well-established utility. As Applicants have asserted utilities that are specific, substantial and credible and well-established, thus the Utility Requirement has been satisfied. Applicants therefore respectfully request the withdrawal of the rejection under 35 U.S.C. § 101.

The Utility Examination Guidelines require a claimed invention to have a utility that is specific to the subject matter claimed (a "specific utility"). The present application recites at, for example, pages 57-58 of the specification that the claimed invention can be used, *inter alia*, to identify ligands and/or protein binding partners. Antibodies can be used to isolate and/or purify the protein from cells to identify other proteins or molecules that interact with the polypeptide to which the antibody specifically binds to. Additionally, the antibodies of the present invention can be used to modulate the activity of the ion channel or related variants (see, for example, page 63). The specification also teaches that antibodies can be used to monitor the expression of the polypeptide in cells or during the development of an organism. Antibodies can also be used to monitor expression of the polypeptide in an *in vitro* setting after transfecting a nucleic acid encoding a polypeptide comprising SEQ ID NO: 105 into cells. Antibodies can also be used to detect modification in the protein in a cell via Western blot. Thus, there is no question that Applicants have asserted at least one specific utility and, in fact, have provided numerous specific utilities for the polypeptides of the present invention.

Additionally, the Office appears to be under the assumption that ***absolute*** certainty is required for a polynucleotide or polypeptide to have a specific utility. The Office states, "There is little doubt that, after complete characterization, this protein will probably be found to have a patentable utility. This further characterization, however, is part of the act of invention and, until it has been undertaken, Applicant's' claimed invention is incomplete." (Office Action, page 5).

The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability as the Supreme Court stated applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner v. Manson*, 383 U.S. at 532. Although, there may be numerous inventions that may arise from the present application, this standard does not justify the Office's stance that the present invention lacks a specific utility. Thus, Applicants have complied with the specific utility requirement.

The Claimed Invention Has A Substantial Utility

The Utility Examination Guidelines also require a claimed invention to have a utility that defines a real-world use (a "substantial utility"). Applicants teach, as described above, that the claimed invention can be used to modulate protein activity, identify ligands and other binding partners, such as other proteins that interact with the polypeptide, monitor expression of the protein *in vivo* or *in vitro*. Thus, it is clear that the claimed invention has real-world uses. All the uses described in the present application are real-world uses and, again, stand in stark contrast to the "throw away" uses (e.g., landfill component or snake food) set forth in the utility guidelines. Thus, there is no question that Applicants have asserted at least one substantial utility and, in fact, have provided numerous substantial utilities. Accordingly, Applicants have complied with the substantial utility requirement.

The Claimed Invention Has A Credible Utility

In addition to a specific and substantial utility, the Utility Examination Guidelines require that such utility be credible (a "credible utility"). That is, whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided. Clearly, the numerous specific and substantial utilities asserted by Applicants are credible. Assertions of credibility are credible unless "(A) the logic underlying the assertion is seriously flawed, or (B) the facts upon which the assertion is based is inconsistent with the logic underlying the assertion." (See, Revised Interim Utility Guidelines Training Materials.) Further, the PTO is reminded that it must treat as true a statement of fact made by Applicants in relation to an asserted utility, unless countervailing evidence can be provided that shows that one of ordinary skill in the art would have a legitimate basis to doubt the credibility of such a statement. All the utilities described for the antibodies and polypeptides are based on sound logic. Furthermore, the utilities for the claimed antibodies are ***not*** inconsistent with the logic underlying the assertion that the antibodies and polypeptides are useful. The Office has provided no evidence that the logic is seriously flawed or that the facts upon which these assertions are based are inconsistent with the logic underlying the assertions.

The Examiner cites literature allegedly identifying difficulties that ***may*** be involved in predicting protein function. None of the cited references, however, suggests that functional homology cannot be inferred by a reasonable probability in any particular case. It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner et al., *Proc. Natl. Acad. Sci.* **95**:6073-78 (1998) (See, attached reference).

In the present application homology is in excess of 40% over many more than 70 amino acid residues. The probability, therefore, that the polypeptide encoded by the claimed polynucleotides is related to the reference polypeptides is, accordingly, very high. None of the references cited by the Examiner contradicts Brenner's basic rule. At most, references cited by the USPTO individually and together stand for the proposition that it may be difficult to make predictions about function with certainty. However, this is not the "countervailing evidence" required by the Utility Examination Guidelines. Therefore, no countervailing evidence that says the present invention does not have a substantial, credible, and useful invention has been provided.

Furthermore, ion channel proteins have a well established utility. Many medically significant biological processes mediated by signal transduction pathways involving ion channels are recognized as important therapeutic targets for a wide range of diseases. In this respect, the ion channel family is analogous to the chemical genus that was the subject of *In re Folkers*, 145 USPQ 390 (CCPA 1965) (Compound that belongs to class of compounds, members of which are recognized as useful, is considered useful under §101.) The Patent Office does not serve the public by attempting to substitute a formulaic analysis of § 101 for the established judgment of the biopharmaceutical industry as to what is "useful." If the Patent Office is aware of any well-grounded scientific literature suggesting that ion channels are not useful, Applicants request that it be made of record.

Art-Recognized Utility

The Utility requirement may also be satisfied by an "Art Established Utility" which means that "a person of ordinary skill in the art would immediately appreciate why

the invention is useful based on the characteristics of the invention... and the utility is specific, substantial and credible." (M.P.E.P. §2107).

Applicants note for the record that the Patent Office apparently agrees with Applicants' reasoning that ion channels are useful in that the Office has granted and apparently continues to grant patents to ion channel proteins, their encoding polynucleotides and antibodies directed to them ***in which no specific biological significance*** is ascribed to the protein. Specifically, Applicants would like to bring the following US Patents to the Office's attention:

6,562,593 Merkulov et al. "Isolated human transporter proteins, nucleic acid molecules encoding human transporter proteins, and uses thereof" (Claims an isolated polynucleotide and method for producing polypeptide)
6,503,733 Bandman et al. "Human anion channel" (Claims an isolated polynucleotide, an isolated polypeptide and an antibody that binds to the polypeptide)
6,228,616 Bandman et al. "Human anion channel" (Claims a purified antibody)
5,854,411 Goli et al. "Human Chloride Channel" (Claims an isolated polynucleotide)
6,451,554 Wood et al. "Ion Channel" (Claims an isolated polynucleotide and a method of producing a polypeptide encoded by the polynucleotide.)
6,309,858 Dietrich et al. "T-type calcium channel variants; compositions thereof; and uses" (Claims an isolated polynucleotide).
6,309,855 Duprat et al. "Family of mammalian potassium channels, their cloning and their use, especially for the screening of drugs" (Claims isolated polynucleotide)
6,207,410 Hall et al. "Genes encoding an insect calcium channel" (Claims isolated polynucleotide and methods)
6,087,488 Ganetzky et al. "Potassium ion channel genes and proteins" (Claims isolated polynucleotide)
6,013,474 Ellis et al. "Calcium channel compositions and methods" (Claims isolated polynucleotide)
5,710,019 Li et al. "Human potassium channel 1 and 2 proteins" (Claims an isolated polypeptide)

Applicants submit that these issued US Patents are evidence of an art recognized utility for ion channels whose natural function or association with disease is unknown. If the Patent Office's position is that issued patents are ***not*** sufficient evidence of art recognition then Applicants respectfully request that this position be made of record. In the alternative, if the Patent Office wishes to take the position that these issued patents are directed to non-statutory subject matter, then Applicants respectfully request that this position be made of record as well.

In view of the foregoing, Applicants respectfully requests that the rejection under 35 U.S.C. § 101 be withdrawn.

Rejections under 35 U.S.C. § 112

Claims 95 and 96 stand rejected under 35 U.S.C. § 112, first paragraph as allegedly failing to adequately teach how to use the instant invention. According to the Office, "Since the claimed invention is not supported by a specific, substantial, and credible asserted utility or a well established utility...one skilled in the art clearly would not know how to used the claimed invention." (Office Action, page 8) Applicants respectfully disagree.

As discussed above, the present invention ***is*** supported by a specific, substantial, and credible asserted utility as well as a well established utility. Accordingly, Applicants respectfully request that the rejection be withdrawn.

Claims 95 and 96 are also rejected under 35 U.S.C. § 112, first paragraph as allegedly containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention. Applicants respectfully disagree

The Office alleges that due to the usage of the word "comprising" the "instant claims encompass virtually every antibody in existence." (Office Action, page 9) Applicants respectfully disagree. One of skill in the art would understand the original claim and that it does not claim "every antibody in existence." Notwithstanding this, in an effort to advance prosecution, Applicants have amended claim 95 to further clarify the scope of the claim.

As presently amended, claim 95 recites, "An isolated antibody which binds to an epitope on a polypeptide comprising SEQ ID NO: 105, wherein said epitope is present within SEQ ID NO:105." A person of ordinary skill in the art would understand that an antibody that is claimed is specific for an epitope within SEQ ID NO: 105 and does not claim an antibody that is already in existence (see, specification, page 51). A person of

ordinary skill in the art would also understand that the antibody would bind to an epitope that is present within SEQ ID NO:105 and not on a peptide or polypeptide sequence that is not present within SEQ ID NO:105.

Additionally, Applicants have added claim 117 that recites, "The isolated antibody of claim 95, wherein said antibody is specific for SEQ ID NO:105." A person of ordinary skill in the art would understand that an antibody that is specific for SEQ ID NO:105 does not cover "virtually every antibody in existence, " rather new claim 117 covers antibodies that are specific for SEQ ID NO:105. Therefore the pending claims are not excessively broad and does not cover "virtually every antibody in existence."

In view of the foregoing, Applicants respectfully request that the rejection of claims 95 and 96 under 35 U.S.C. § 112, first paragraph be withdrawn.

Rejections under 35 U.S.C. § 102

Claims 95 and 96 stand rejected under 35 U.S.C. § 102(b) as allegedly anticipated by Hopp *et al.* (U.S. Patent No. 5,011,912, hereinafter "Hopp"). Applicants respectfully disagree.

The standard for anticipation under 35 U.S.C. § 102 is one of strict identity. An anticipation rejection requires a showing that each limitation of a claim be found in a single reference, *Atlas Powder Co. v. E.I. DuPont de Nemours & Co.*, 224 U.S.P.Q. 409, 411 (Fed. Cir. 1984).

The Office alleges that:

> [c]laims 95 and 96 are directed to an antibody which binds to an epitope on a polypeptide encoded by an isolated nucleic acids molecule comprising a nucleotide sequence that encodes a polypeptide comprising an amino acid sequence homologous to SEQ ID NO:105. Because of the use of "comprising" language in the preceding claims, the instant claims encompass virtually every antibody in existence. Hopp et al. disclose a monoclonal antibody that meets the limitation of the instant claims.

(Office Action, page 11, emphasis in original). As amended, claim 95 recites an isolated antibody which binds to an epitope on a polypeptide comprising SEQ ID NO: 105, wherein said epitope is present within SEQ ID NO:105. Hopp does not discuss or even

suggest an antibody that binds to an epitope present within SEQ ID NO:105. Therefore, Hopp fails to anticipate claims 95 and 96.

In view of the foregoing, Applicants respectfully request that the rejection of claims 95 and 96 under 35 U.S.C. § 102 (b) be withdrawn.

Rejections under 35 U.S.C. § 103

Claims 95 and 96 stand rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over Isenberg *et al* (Neuroreport, 1993, 5 pp. 121-124, hereinafter "Isenberg"). According to the Office, Isenberg discusses an amino acid sequence that comprises "an epitope of eight consequent amino acids, which completely match an epitope of SEQ ID No:105...Therefore an antibody which binds to an epitope of a fragment of 5HT3 receptor of Isenberg et al. would anticipate the instant claims." (Office Action, page 12). As the Office admits, Isenberg does not disclose an antibody that binds to a fragment of the 5HT3 receptor. Applicants respectfully point out that even if Isenberg did disclose an antibody that binds to the receptor, it would have to be demonstrated that the antibody would also bind to an epitope within SEQ ID NO: 105 for it to anticipate the claimed invention.

The Office further states, "At the time the invention was made it would have been *prima facie* obvious to one of ordinary skill in the art to generate antibodies to 5HT3 receptor for purposes of tissue printing and localization of 5HT3. Such antibodies would be encompassed by claims 95 and 96." (Office Action, page 12). Applicants respectfully disagree.

In establishing a *prima facie* case of obviousness under 35 U.S.C. §103, it is incumbent upon the Examiner to provide a reason why one of ordinary skill in the art would have been led to modify a prior art reference or to combine reference teachings to arrive at the claimed invention. *Ex parte Clapp*, 227 U.S.P.Q. 972 (Bd. Pat. App. Int. 1985). To this end, the requisite motivation must stem from some teaching, suggestion or inference in the prior art as a whole or from the knowledge generally available to one of ordinary skill in the art and not from appellants' disclosure, see for example, *Uniroyal*

Inc. v. Rudkin-Wiley Corp., 5 U.S.P.Q.2d 1434 (Fed. Cir. 1988); and *Ex parte Nesbit*, 25 U.S.P.Q.2d 1817, 1819 (Bd. Pat. App. Int. 1992). In this respect, the following quotation from *Ex parte Levengood*, 28 U.S.P.Q.2d 1300, 1302 (Pat. Off. Bd. App. 1993), is noteworthy:

> Our reviewing courts have often advised the Patent and Trademark Office that it can satisfy the burden of establishing a *prima facie* case of obviousness only by showing some objective teaching in either the prior art, or knowledge generally available to one of ordinary skill in the art, that "would lead" that individual "to combine the relevant teachings of the references." ... Accordingly, an examiner cannot establish obviousness by locating references which describe various aspects of a patent applicant's invention without also providing evidence of the motivating force that would **impel** one skilled in the art to do what the patent applicant has done. (citations omitted; emphasis added)

Significantly, the Office Action identifies no "motivating force" that would "impel" persons of ordinary skill to modify the respective teachings of the cited reference and achieve the claimed invention. The only motivation for modifying the reference in the manner identified in the Office Action is the general statement, "it would have been...obvious," nothing that would have impelled a person of ordinary skill in the art to modify the Isenberg reference.

Furthermore, there is nothing in the Isenberg reference that would motivate a person of ordinary skill in the art to generate an antibody to the eight amino acid sequence that was identified by the Office as being identical to an eight amino acid sequence in SEQ ID NO:105. The eight amino acids are a sub-part of a polypeptide that is 464 amino acid residues in length. It appears that the Office has chosen those eight residues based solely on the Applicants' disclosure, which is strictly prohibited.

Thus, the Office has provided no motivation that would impel one of skill in the art to even select an eight amino acid epitope from the unrelated protein described in Isenberg, less still develop an antibody to the eight amino acid epitope and show binding. Therefore, Applicants respectfully assert that the claimed invention is ***not*** obvious.

In view of the foregoing, Applicants respectfully request that the rejection under 35 U.S.C. § 103(a) be withdrawn.

Conclusion

Applicants believe the claims are in condition for allowance. An early Notice of Allowance is therefore earnestly solicited. Applicants invite the Examiner to contact the undersigned at (215) 665-6928 to clarify any unresolved issues raised by this response.

Respectfully submitted,

Daniel M. Scolnick, Ph.D.
Reg. No. 52,201

Date: July 29, 2003
COZEN O'CONNOR, P.C.
1900 Market Street
Philadelphia, PA 19103-3508
Telephone: (215) 665-2000
Facsimile: (215) 665-2013

Attachment: Brenner *et al. Proc. Natl. Acad. Sci.* **95**:6073-78 (1998).

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "lnscaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

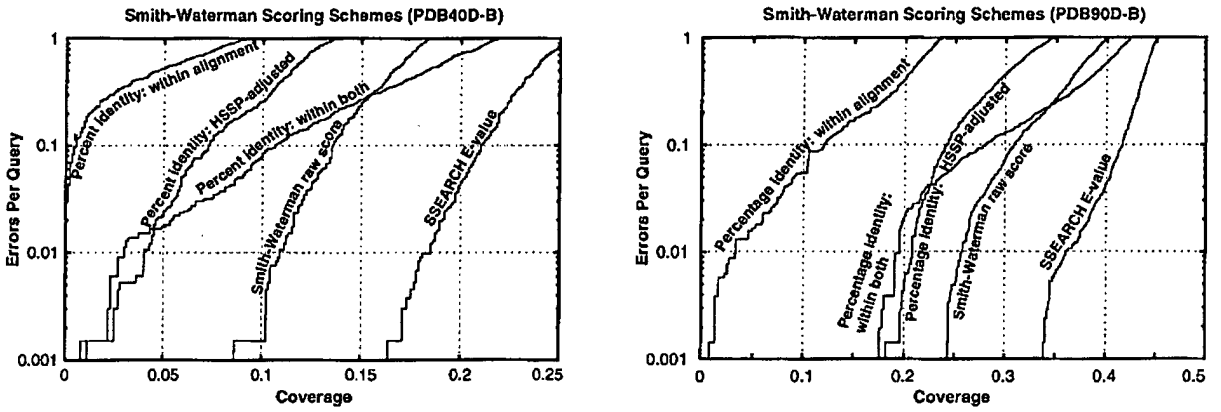


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

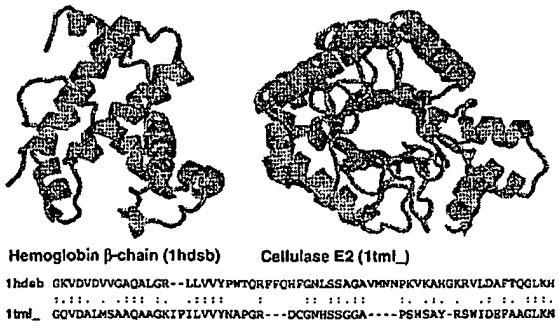


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

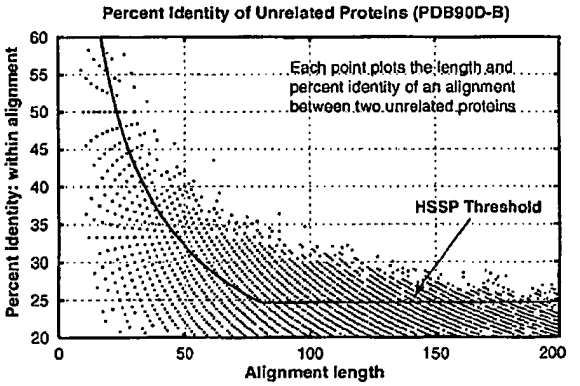


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

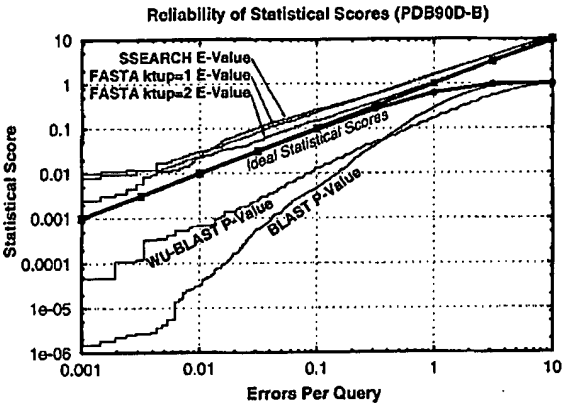


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

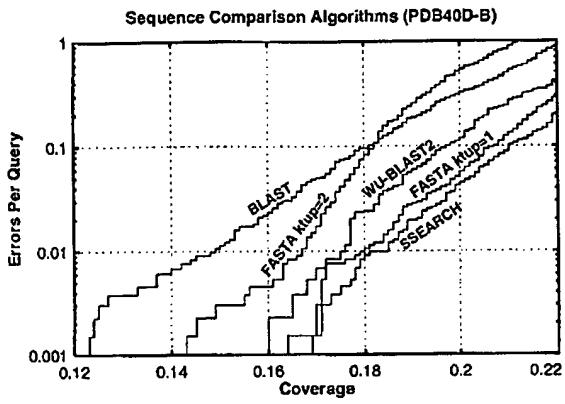
Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most



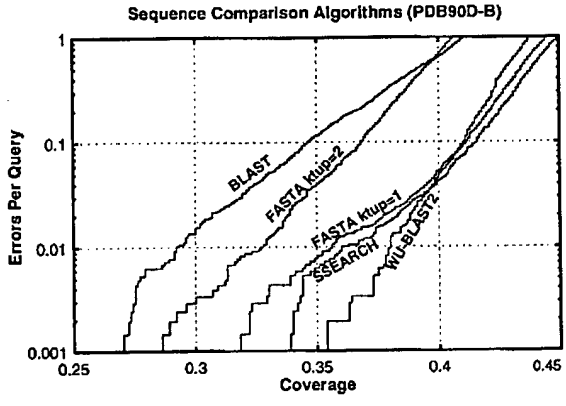


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

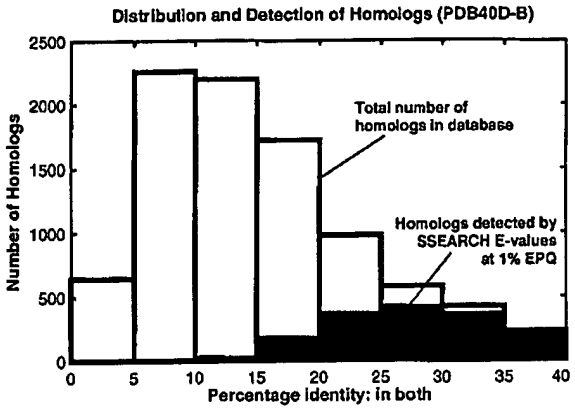


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.